DOCUMENT RESUME

ED 211 606 TM 820 062

AUTHOR Fuchs, Lynn; And Others
TITLE Effects of Varying Item Domain and Sample Duration on Technical Characteristics of Daily Measures in Reading.
INSTITUTION Minnesota Univ., Minneapolis. Inst. for Research on Learning Disabilities.
SPONS AGENCY Office of Special Education (ED), Washington, D.C.
REPORT NO IRLD-RR-48
PUB DATE Jan 81
CONTRACT 300-80-0622
NOTE 44p.
AVAILABLE FROM Editor, IRLD, 350 Elliott Hall, 75 East River Road, University of Minnesota, Minneapolis, MN 55455 ($3.00)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Education; *Item Banks; Learning Disabilities; *Reading Ability; Reading Tests; *Sampling; Test Items; *Test Reliability; *Word Recognition
IDENTIFIERS *Sample Size; Test Curriculum Overlap

ABSTRACT

Three related studies were conducted to examine the effects of variations in procedures used for curriculum-based assessment of reading proficiency: the first addressed the question of the influence of sample duration on the concurrent validity of the measure; the second addressed the question of the influence of sample duration on the level, slope, and variability of performance over repeated measurements; and the third examined the effect that varying the size of the pool from which items are drawn has on slope and variability of performance on the measure. Results of the studies provided evidence that sample duration is an important consideration in curriculum-based measurement because of its probable impact on variability and slope. Increasing sample duration from 30 seconds to a three minute sample reduced day-to-day variability in performance and resulted in a more rapid increase in student performance. The results with respect to sampling from domains of differing sizes indicated that measurement samples drawn from smaller domains are more sensitive to variations in instruction, but somewhat more variable. The optimum daily measurement procedure would seem to involve sampling from a pool of stimulus items well beyond that defined by the short-term objectives, but not in excess of an annual goal. (Author/GK)

University of Minnesota

Research Report No. 48

EFFECTS OF VARYING ITEM DOMAIN AND SAMPLE DURATION ON TECHNICAL CHARACTERISTICS OF DAILY MEASURES IN READING

Lynn Fuchs, Gerald Tindal, and Stanley Deno

IRLD

Institute for Research on Learning Disabilities

IRLD

Director: James E. Ysseldyke

Associate Director: Phyllis K. Mirkin

The Institute for Research on Learning Disabilities is supported by a contract (OEG-80-0622) with the Office of Special Education, Department of Education through Title VI-G of Public Law 91-230. Institute investigators are conducting research on the assessment/decision-making/intervention process as it relates to learning disabled students.

During 1980-1983, Institute research focuses on four major areas:

- Referral
- Identification/Classification
- Intervention Planning and Progress Evaluation
- Outcome Evaluation

Additional information on the Institute's research objectives and activities may be obtained by writing to the Editor at the Institute (see Publications list for address).

The research reported herein was conducted under government sponsorship. Contractors are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent the official position of the Office of Special Education.

Research Report No. 48

EFFECTS OF VARYING ITEM DOMAIN AND SAMPLE DURATION ON TECHNICAL CHARACTERISTICS OF DAILY MEASURES IN READING

Lynn Fuchs, Gerald Tindal, and Stanley Deno

Institute for Research on Learning Disabilities

University of Minnesota

January, 1981

Abstract

Three related studies were conducted to examine the effects of variations in procedures used for curriculum-based assessment of reading proficiency. The first study addressed the question of the influence of sample duration on the concurrent validity of the measure. The second study addressed the question of the influence of sample duration on the level, slope, and variability of performance over repeated measurements. The third study was designed to examine the effect that varying the size of the pool from which items are drawn has on slope and variability of performance on the measure.

The results of the three studies provided evidence that sample duration is an important consideration in curriculum-based measurement because of its probable impact on variability and slope. Increasing sample duration from 30 seconds to a three minute sample reduced day-to-day variability in performance and resulted in a more rapid increase in student performance. The results with respect to sampling from domains of differing sizes indicated that measurement samples drawn from smaller domains are more sensitive to variations in instruction, but somewhat more variable. The optimum daily measurement procedure would seem to involve sampling from a pool of stimulus items well beyond that defined by the short-term objectives, but not in excess of an annual goal.

Effects of Varying Item Domain and Sample Duration on Technical Characteristics of Daily Measures in Reading

As the limitations of standardized testing for use in instructional programming become clearer, interest has increased in using routine measurement of student performance on curriculum objectives as the basis for improving educational decisions (Jenkins, Deno, & Mirkin, 1980; Lovitt, 1977; Popham, 1980). Evidence has begun to accumulate that, indeed, instructional effectiveness can be increased by having teachers measure student performance and use those data to set goals and evaluate changes in methods and materials (Bohannon, 1975; Crutcher & Hofmeister, 1972; Frumess, 1973; Lovitt, Schaff, & Sayre, 1970; Mirkin & Deno, 1979; Mirkin, Deno, Tindal, & Kuehnle, 1980).

The logical and empirical arguments for increased emphasis on using frequent measurement of student performance on curriculum objectives has been accompanied by the concurrent development of training materials designed to teach teachers how to do such measurement (Deno & Mirkin, 1977; Howell, Kaplan, & O'Connell, 1979; White & Haring, 1976). Further, a substantial number of demonstration projects have been funded that include as a major component the use of curriculum-based daily measurement (NaLDAP, 1976; NaLDAP, 1978; PDAS, 1980).

As momentum gathers for using curriculum-based assessment to make instructional programming decisions, concern increases for precisely how to do such measurement. In contrast to standardized testing where test items and procedures are made available to the consumer, curriculum-based testing requires the teacher to create continuously the test

materials and procedures for use with individual students. While the technical characteristics of many commercially published standardized tests are known, little is known regarding the technical characteristics of curriculum-based testing. Since variation in test procedures has a significant bearing on the reliability and validity of standardized tests, we should examine the effects of variations in procedures for measuring student performance on curriculum objectives.

The purpose of this paper is to report on three related studies conducted to examine the effects of variations in procedures for curriculum-based assessment of reading proficiency. The first study addressed the question of the influence of sample duration on the concurrent validity of the measure. The second study addressed the question of the influence of sample duration on the level, slope, and variability of performance over repeated measurements. The third study was designed to examine the effect that varying the size of the pool from which items are drawn has on slope and variability of performance on the measure.

STUDY-I

Research has demonstrated that one minute word recognition measures correlate highly with reading comprehension measures as well as with standardized reading tests (Deno, Mirkin, Chiang, & Lowry, 1980). A simple word recognition test, therefore, appears to be a valid index of a student's reading proficiency. Given its ease of administration and the availability of alternate forms, a simple word recognition measure might be employed as a measure for monitoring reading progress.

Several issues related to the parameters of test construction need to be addressed, however. Variations in measurement procedures, such as shortening test duration, increase test efficiency and render a word recognition test more practical as a formative evaluation measure. At the same time, variations in measurement procedures can affect a test's technical adequacy.

Study I was designed to examine how the duration of a curriculum-based test sample affects two dimensions of a measure's technical adequacy, specifically: (a) concurrent validity, and (b) variability of performance.

Method

Subjects. Twenty-seven (M=17, F=10) students were randomly selected from grades 1-6 in two Minneapolis public elementary schools. In addition, 18 (M=13, F=5) students were recruited from the learning disability resource programs in those two schools.

Materials. Five curriculum-based measures (Words in Isolation, Words in Context, Oral Reading, Cloze Comprehension, and Word Meaning) whose criterion validity had already been determined were employed. To be included, a measure had to have potential for routine use by classroom teachers.

The Words in Isolation measure consisted of four alternate forms of randomly selected words from the Core List of 5,167 words listed in Basic Elementary Reading Vocabulary - R Series (Harris & Jacobson, 1972). Two lists were samples from each of Pre-Primer through third grade levels and two lists were samples from Pre-Primer through sixth grade. Words were included on the word lists only if they had a frequency index of more than 10 per million words in the Teacher's Word Book of 10,000 Words (Thorndike & Lorge, 1944).

The Words in Context measure consisted of passages of approximately 600 words selected from the beginning, the middle, and the latter parts of books for three different basal reading series: Allyn-Bacon, Ginn 720, and Houghton-Mifflin. Two passages sampled from sixth grade books and two sampled from third grade books. Words were typed with every fifth word underlined in each passage (see Appendix B in Deno et al., 1980). The reading levels for these passages were computed using the Fry Readability Index formula (Fry, 1968), and each passage was at the appropriate difficulty level, either third or sixth grade.

The Oral Reading measure included four passages of 300 words each. These were selected from the basal readers and typed on sheets of paper (see Appendix C in Deno et al., 1980). The reading levels for the passages were again computed using the Fry Readability Index formula (Fry, 1968) and each was at the appropriate level.

The Cloze measure was developed from four additional passages of 300 words each that were selected from the same basal readers. The first and last sentence in each passage was left intact, but every tenth word was deleted from all other sentences in the passage. The passages were then typed with five-space blanks in place of the deleted words (see Appendix D in Deno et al., 1980).

The Word Meaning measures involved the use of three passages consisting of 300 words each that were selected from the same basal readers. Every fifth word of the passage that was clearly definable and not a function word (i.e., an article, preposition, proper noun) was underlined (see Appendix E in Deno et al., 1980).

Procedure. The five measures were individually administered in one session. Each subject was taken to a quiet room by a psychometrician who had been trained to administer and score these measures. Each student was given the measures in the following order: Words in Isolation, Words in Context, Oral Reading, Cloze, Word Meaning. The students completed two 30-second and two 60-second tests on parallel forms for each of the word recognition measures. For the Cloze measure, each test was two minutes.

The Words in Isolation test instructions were read verbatim to the subject:

> Here is a word list that I want you to read. When I tell you to start, you can read across the page. Use the cardboard to help you keep your place. Please read as fast and accurately as you can. If you get stuck on any of the words, move on to the next one. I will tell you when to stop reading. Are there any questions? Ready? Begin.

Then the word list was given to the child and the stopwatch was triggered for the appropriate duration. A psychometrician marked whether each word was correctly read on a follow-along sheet that was identical to the word list itself. If the child failed to respond after an interval of approximately six seconds, the psychometrician urged the child to move on to the next word. Immediately following the timing of the first word list, the remaining lists were administered consecutively. Responses had to be completely accurate to be scored as correct.

The procedures for Words in Context were similar to those used for Words in Isolation. The following instructions were read to the child:

> I am going to show you a story that has underlined words in it. Say the underlined words as quickly and accurately as you can. Start at the top of the page and try not to skip any words. If you do not know a word, try the next word. Here is a cardboard strip that you can use to help you keep your place. Remember to do the best that you can, and I will tell you when the time is up. Are you ready? Here is the story. Begin.

The four lists were given one after the other, each for the appropriate sample duration. Words had to be read accurately to be scored as correct.

The Oral Reading passages were read during consecutive timings after the following instructions were given:

> Now I am going to give you a story that I would like you to read aloud to me. Do your best and go on reading if you get stuck on a word. I'll let you know when to start and stop. Do you have any questions? Remember to do your best, but do not take a lot of time on hard words. Here's the story. Ready? Begin.

Omissions, insertions, substitutions, and mispronunciations all were tallied as errors.

The Cloze passages were then administered. A sample passage was included at the beginning of the first cloze passage to ensure that subjects understood the task. The test instructions were:

> I'm going to give you a story that has some words missing in it. You are to try to read the story and fill in the blanks of the missing words. Let's read the first sentence together. [Sentence read with subject.] Now read aloud the next sentence and try to fill in the blank of the missing word. [Subject reads sentence.] It is not easy to guess what the missing word could be, but do the best you can. If you cannot put a word into the blank, move on to the next blank and try to work quickly. Are you ready to begin? Begin.

Synonyms of the deleted words were considered correct.

The instructions for the Word Meaning measure were:

> I am going to show you a story that has underlined words in it. Tell me the meaning of the underlined words. Try to do your best and work quickly. If you do not know the meaning of a word, skip it, and go on to the next word. You can use the cardboard strip again to help you keep your place. Remember to do your best, and I will tell you when the time is up. Are you ready? Here is the story. Look at the first line and tell me the meaning of the underlined words. Begin.

Psychometricians had been trained on the types of responses that were acceptable. Decisions were made regarding the correctness of each response immediately after the response was given.

Results

Concurrent validity. For purposes of analysis, a mean score was computed using the pairs of 30-second and 60-second scores for each student. The data, then, consisted of: 12 word recognition scores (2 levels X 2 test times X 3 types of word recognition), one cloze score, and one word meaning score. The descriptive data for these scores, including group means and standard deviations, appear in Table 1.

Insert Table 1 about here

The 14 scores for each student were then intercorrelated. Tables 2, 3, and 4 contain the correlation matrices for the 14 variables from the resource, regular, and combined groups, respectively. The median correlation for the combined groups between the 30-second and 60-second samples was .92, with a range of .83 to .97. The median correlation between the 30-second sample and the Cloze measure was .86, with a range of .76 to .86; the median correlation between the short sample and Word Meaning was .61, ranging from .49 to .71. All correlations were statistically significant ($p < .001$).

Insert Tables 2-4 about here

Variability. A standard deviation was calculated for the group scores on each 30-second and 60-second measure (see Table 1). These standard deviations were then averaged across the 30-second measures and across the 60-second measures. The mean standard deviation for the 30-second samples was 14.12; the mean standard deviation for the 60-second

samples was 27.60. The discrepancy between these average values was subjected to a correlated t test, which revealed a statistically significant difference ($p < .001$).

Discussion

The 30-second and 60-second samples consistently correlated very highly with each other. The 30-second samples and the reading comprehension tests also correlated significantly and always similarly to the way that the 60-second samples and reading comprehension measures correlated. This study, therefore, directly demonstrates the concurrent validity among 30-second and 60-second samples of word recognition measures and reading comprehension measures. Additionally, because the 60-second word recognition measures employed in this study had previously demonstrated consistently high correlations with standardized reading tests (Deno et al., 1980), Study I indirectly establishes concurrent validity among 30-second word recognition measures and standardized reading tests. Also, the lower average standard deviation for the 30-second samples as compared to the 60-second samples indicates that these shorter tests result in reduced variability and improved reliability. This demonstrates that shorter durations may improve the technical adequacy of simple, direct measures. On the basis of the results of Study I, therefore, one can conclude that the 30-second duration samples, which are logistically more feasible forms of the word recognition tests, are as valid and reliable indices of reading proficiency as the 60-second samples.

STUDY II

Employing group data, Study I confirmed two dimensions of the

technical adequacy of 30-second word recognition tests by demonstrating their concurrent validity with 60-second word recognition and reading comprehension measures and by revealing that, across groups, they result in reduced variability. Unfortunately, measurement theory (Kelley, 1927; Nunnally, 1959) warns that apparently adequate technical data may have limited applicability to individual assessment. The standard error of the group performance may substantially reduce the relevance of group technical data in the interpretation of individual scores. Therefore, in examining the technical adequacy of formative measurement instruments that are employed to test individual performance only, it is important to investigate measurement issues that directly relate to the reliability and validity of time series data.

One characteristic of technically adequate time-series measurement instruments is that they result in low variability in the data. Reduced variability is important, because as variability between data points decreases, the reliability of the measure increases, the relative effectiveness of different phases in formative evaluation is more easily and quickly determined, and any one data point provides more information about a student's true score.

As one judges the technical adequacy of a measurement format by investigating its influence on the variability in the data, one must simultaneously examine that format's effect on the level and slope of a student's performance. In fact, evidence suggests that characteristics of the measurement procedure itself may not only influence the variability of the data but also affect the rate and trend of a student's performance (Ayllon, Garber, & Pisor, 1976).

As in Study I, the purpose of this experiment was to examine the influence of variations in sample duration on the technical adequacy of a simple word recognition measure. In contrast to the first study, however, Study II assessed technical adequacy by employing a single case experimental design, simultaneously examining the relationship among duration of measurement sample and the level, slope, and variability of time series data.

Method

Subjects and setting. The students who served as subjects in the study had been designated as reading "seriously" below their teachers' expectations during a Title I needs assessment. As a result they were enrolled in a Title I reading room program that provided daily, supplementary help to students in their regular classroom basal readers. This program serviced approximately 40% of the kindergarten through third grade student body of an inner city midwestern metropolitan school.

The study included two children who were selected because of their consistent school attendance and because of their similarity to each other. These two second grade, eight-year-old girls shared a classroom; they were grouped together in level five of the Ginn 720 readers; both worked on the same phonics categories within the Title I reading program; and, over a five-week interval, both consistently scored within five words of each other on weekly, one-minute samples of the number of correct consonant-vowel-consonant patterned words read from flashcards.

Procedure. The experimental questions were examined through the use of a combined multiple baseline across subjects and reversal design (Hersen & Barlow, 1976), consisting of four experimental phases: Phase A, a daily

30-second measurement sample: Phase B, a daily three-minute measurement sample; Phase C, return to a daily 30-second measurement sample; and Phase D, return to a daily three-minute sample.

Student 1 began Phase A; after six days, this student entered Phase B and Student 2 simultaneously began Phase A. Similarly, phases were allowed to run five to nine days before the students progressed to their next phases. Throughout the experiment, the dependent data were the number of correctly read consonant-vowel-consonant patterned words per minute and the number of incorrectly read consonant-vowel-consonant patterned words per minute. The Title I reading teacher individually collected the data at the end of the students' standard 20-minute instructional session. With a stopwatch and a shuffled 3x5 inch deck of consonant-vowel-consonant patterned word cards, she exposed each card for a maximum of two seconds to the student and then placed the card into a correct or incorrect pile. When the allotted time expired, the teacher counted words correct and words incorrect and recorded the scores on a form provided by the experimenter.

Results

Level of student performance. The dependent data for both students are shown in Figure 1. An analysis of this graph reveals that the median level performance of words correct was consistently higher in the 30-second presentations than in the three-minute presentations. Despite this superior level of performance, however, in three out of the four 30-second phases, the trends are flat, while the trends in all four three-minute phases are accelerating. The consistently higher median performances in the 30-second phases appear to be related to the initial step down with each introduction of a three-minute phase.

Insert Figure 1 about here

Variability. The variability of each phase was summarized in two different ways. First, the total bounce (Pennypacker, Koenig, & Lindsley, 1972) was calculated. Total bounce (TB) is the distance between the line parallel to the trend line passing through the frequency dot farthest above the trend line and the line parallel to the trend line passing through the frequency dot farthest below the trend line. The solid and dotted lines in Figure 2 display the TB for each phase of the experiment: Table 5 presents the TB scores for each phase as well as the average of the 30-second and three-minute phases and the grand average of all 30-second phases and of all three-minute phases.

Insert Figure 2 and Table 5 about here

The second method employed to summarize variability was the standard error of the estimate (SEE) of the trend line. This SEE is calculated for each phase by taking the square root of the average of the squared deviations of each point from the trend line, which was determined using the split-median solution (White, 1971). Table 5 presents the SEE for each phase and for the average of 30-second and three-minute phases for each student and the grand average of all 30-second phases and of all three-minute phases.

By inspecting the TBs in Figure 2, one can readily see that the 30-second phases were more variable than the three-minute phases. Moreover, Mann-Whitney tests on the TB and SEE scores revealed statistically significant differences in the variability between the 30-second and three-minute

phases (two tailed p = .037 and .043, respectively).

Discussion

An analysis of the relationship between the level of performance and the duration of measurement sample yielded conflicting results. The median levels of performance for the 30-second phases were consistently higher than the levels of performance for the three-minute phases, a comparison that demonstrated the superiority of the 30-second presentations. The analysis of the trends within the phases, however, showed accelerating trends in all of the longer presentation phases and flat slopes in three out of four shorter presentation phases. It is possible that given longer phases for the three-minute presentations, performance under the longer measurement condition might surpass performance under the 30-second presentations. Therefore, although the duration of measurement condition exhibited a consistent controlling effect, the exact nature of that effect is unclear and the superiority of one sample duration over the other is not evidenced.

The most dramatic result in this study was the greater variability for the 30-second phases compared to the three-minute phases. As stated above, a decrease in variability directly relates to the concerns of both practitioner and researcher, because with reduced variability, reliability of the measurement improves, stability of the data increases, the relative effectiveness of different phases in single case research designs is more easily determined, and any one data point provides more information about a student's true score.

Additionally, O'Connor and Weiss (1974) suggest that a measure's validity also increases as variability decreases. However, to anticipate such an increase in validity, one must assume that the altered method of

administration of the measure, here the prolonged sample duration, does not alter the abilities tapped by that measure. In this experiment, it is possible that the prolonged timings do not yield more valid data, but rather reflect a change in the nature of what is being measured. The three-minute presentations might measure, in addition to reading skills, the students' concentration skills. The initial drop in performance level with each introduction of a three-minute phase and the subsequent accelerating trend, then, might be explained by the students' initially poor concentration over the prolonged sample, which improved with practice over the phase. Within this scheme, one might hypothesize that given longer runs of the three-minute timings, the accelerating trend might level off as concentration approaches a ceiling level for the students.

Study II, then, revealed that a longer sample duration resulted in reduced intra-individual variability and increased reliability of the time series data. In contradistinction, Study I revealed that shorter samples produced lower inter-individual variability.

The results of Studies I and II, therefore, seem contradictory and confusing. Yet, tests should be validated within the context in which they will be used (Cronbach, 1971). Given the purpose of simple, direct measures to evaluate behavior on an on-going basis, the time series analyses of variability performed in Study II appear to bear more directly on the technical adequacy of simple and direct measures. The results of these analyses tentatively support the use of longer measurement durations.

Study III

In Studies I and II test duration was examined because reducing

test time increases teacher efficiency. A second issue to be addressed in measuring word recognition performance is the size of the domain from which test words are drawn. Domain size is an important factor because of its potential impact on teachers' data utilization. Data on samples from larger domains provide teachers with a basis for broader generalizations about performance than do data sampled from more limited domains. The differences in performance might lead to different program evaluation decisions. Samples from smaller domains constitute more direct measures of performance (Lovitt, Schaff, & Sayre, 1970), and may provide teachers with more immediate feedback on the effectiveness of instructional interventions. Larger domains provide teachers with richer data on progress towards long-term goals. Additionally, a large domain is preferable because, once established as the pool from which repeated measures are drawn, it can remain intact and provide comparable data over an extended time.

As in the case of test duration, domain size might well impact the technical characteristics of the test data. Therefore, in Study III, the effect of the domain size on the slope and variability of student performance was investigated.

Method

Subjects. Twenty students in a metropolitan school district, reading at the second, third, or fourth grade instructional levels, served as subjects.

Materials. Reading measures were developed, each of which was a list of 60 core words appearing in Basic Elementary Reading Vocabulary - R Series (Harris & Jacobson, 1972), a compilation of over 500 words used

in several basal readers. Twenty-five lists were generated from each of the following domains: 1) in the most limited domain, the grade-specific domain (GS), 200 words were randomly selected from each grade level. The 25 different word lists were developed by randomly sampling from this domain of 200 words; 2) in a more comprehensive domain, words from the entire grade level (GE) provided the pool of words from which 25 different word lists were devised; 3) the largest domain, across-grade domain (AG), consisted of words from the entire pool of words appearing in preprimer through grade 4, with the 25 different word lists sampled from across these grades.

Procedure. For the first five days teachers placed each student for reading instruction employing the following procedure. The student read from each of the GE lists (preprimer-grade 4) for 30 seconds and the teacher recorded the number of words read correct and incorrect for each of the four word lists. The student was placed for instruction at the grade level in which the median number of words read correct was the highest.

Beginning the second week (6th day), the teachers began instructional programs for all their students, using the words from the 200 word list (GS) representing the student's instructional level. Each student was individually instructed for ten minutes daily. Immediately following each instructional period, the teacher administered three 30-second tests; one from the appropriate grade level (GE), one from the appropriate instructional level (GS), and one from the across-grade domain (AG).

Results

To determine the effects of sampling from domains of different sizes on the slope of student performance and the variability around the slope,

the mean slope and the mean standard error of the estimate (SEE) were computed for the data generated from each domain. Table 6 presents the average slope and SEE for each of the three domains. The means for slope and SEE were then compared using t tests for correlated data and the results of these comparisons are presented in Table 7. As can be seen in Table 7, a statistically significant difference in the slope was obtained between GS and GE data. At the same time the SEE on the data from these two domains revealed that the variability around the slope was significantly greater with the data from the GS words than that for the GE words. The same analysis for the contrast between GE data and AG data revealed a reliable difference in the slope, but no difference in the SEE. When samples were drawn from the AG domain student performance resulted in a nearly flat slope (-.07).

Insert Tables 6 and 7 about here

Discussion

The results provide evidence that when measuring student performance in reading isolated words on a daily basis, the average slope of student performance is likely to decrease as the size of the domain from which the samples are drawn increases. The slope was steepest when the sampling procedure was limited to a 200 word subsample from the grade level at which the student was being instructed. There was a decrease in the slope when the domain was all the words from the grade level. Finally, the slope fell to near zero when the domain spanned several grades. While there were consistent differences in the slopes for the three domains, the differences in the SEE were inconsistent. The degree of variability for the largest and

smallest domains was similar, with less variability in performance on the intermediate domain. Since it is difficult to generate a plausible hypothesis accounting for this result, the obtained effect may well be an artifact of the procedures used. It is important to determine which data are misleading.

From the standpoint of routine measurement, it would seem that the most useful procedures would be those producing time series-data with steep slopes and minimal variability. A steep slope indicates rapid growth and provides a scale that can be sensitive to short term treatments. Similarly, procedures that result in low variability provide more precise estimates of both level and slope of performance, thereby increasing the reliability of conclusions about the effects of changes in an instructional program. The present results indicate that a domain somewhat beyond that defined by the short term instructional objectives might be the best choice for sample selection. The slope of performance based on sample words drawn from the entire reading vocabulary for the student's grade level is likely to be sufficiently steep and at the same time the standard error for that same data should be relatively small. In terms of IEP goals, the results provide technical support for repeatedly measuring performance on the annual goal as a means of generating data for continuously evaluating program success. Nevertheless, measurement based on the immediate objectives of instruction - as is the case when measuring from the current week's reading vocabulary - appeals to the practitioner since it provides evidence of whether the student is learning what currently is being taught. Further, present results also indicate that daily performance gains on smaller domains are substantially greater. If the greater variability in performance

obtained for the smallest domain is indeed an artifact, drawing measurement samples from more immediate instructional objectives may well be preferable.

Conclusions

Taken together, the results of the three studies reported here provide an empirical basis for two major conclusions regarding the procedures used to measure repeatedly reading performance in the curriculum. First, while varying the test duration from one-half to one minute has little impact on the criterion validity of the isolated word recognition task, increasing test duration from one-half minute to three minutes substantially reduces variability in repeated testing over time. Therefore, if reading performance is measured by repeated reading of isolated words for one-half to one minute and it is difficult to estimate the level and trend of performance because of high variability from test to test, then a more precise performance estimate can be attained by increasing test time.

A second conclusion to be drawn from the research relates to the domain from which test stimuli should be drawn. Previous related research (Deno, Mirkin, Chiang, & Lowry, 1980) has provided evidence that, within limits, the difficulty level of the words used as test stimuli has little influence on the test's power to discriminate high from low achievement in reading. In the present research, however, evidence was obtained that test stimuli drawn from smaller domains at the student's instructional level might be more useful for evaluating the effects of instruction. At the same time, testing from smaller domains at instructional level will force frequent changes in the population of test stimuli as a

student achieves mastery in the domain. If domains must be changed, then changes in the level and slope of a student's performance will be a function of changes in the measurement system. Since the purpose of frequent repeated measurement essentially is to determine whether a program is effectively increasing achievement and whether adjustments in a program are having the intended effect, changing the measurement system by introducing new stimulus items will make it difficult to draw valid conclusions regarding program effects. Results from the present study provide support for drawing test stimuli from a domain defined by estimating what a student might be expected to attain in one-half to one full school year. To draw samples from smaller domains will result in the need to change frequently the measurement domain. To draw measurement samples from domains defined by goals exceeding what can be attained by the student within one year is likely to result in a measurement system insensitive to program adjustments.

References

Ayllon, T., Garber, S., & Pisor, K. Reducing time limits: A means to increase behavior of retardates. Journal of Applied Behavior Analysis, 1976, 9, 247-252.

Bohannon, R. Direct and daily measurement procedures in the identification and treatment of reading behaviors of children in special education. Unpublished doctoral disseration, University of Washington, Seattle, 1975.

Cronbach, L. J. Test validity. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Crutcher, C. E., & Hofmeister, A. M. Effective use of objectives and monitoring. Teaching Exceptional Children, 1975, 7(2), 78-80.

Deno, S., & Mirkin, P. Data-based program modification: A manual. Minneapolis: Leadership Training Institute/Special Education, University of Minnesota, 1977.

Deno, S., Mirkin, P., Chiang, B., & Lowry, L. Relationships among simple measures of reading and performance on standardized achievement tests (Research Report No. 20). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities, 1980.

Frumess, S. C. A comparison of management groups involving the use of the standard behavior chart. Unpublished doctoral dissertation, 1973.

Fry, E. Graph for estimating readability. Journal of Reading, 1968, 577.

Harris, A. P., & Jacobsen, M. D. Basic elementary reading vocabularies. New York: Macmillan, 1972.

Hersen, M., & Barlow, D. H. Single case experimental designs: Strategies for studying change. New York: Pergammon Press, 1976.

Howell, K., Kaplan, J., & O'Connell, C. Y. Evaluating exceptional children: A task analysis approach. Columbus, Ohio: Charles Merrill, 1979.

Jenkins, J. R., Deno, S. L., & Mirkin, P. Measuring pupil progress toward the least restrictive alternative. Learning Disability Quarterly, 1979, 2, 81-91.

Kelley, T. L. Interpretation of educational measurements. Yonkers-on-Hudson, N.Y.: World Book, 1927.

Lovitt, T. C. In spite of my resistance, I've learned from children. Columbus, Ohio: Charles E. Merrill, 1977.

Lovitt, T., Schaff, M., & Sayre, E. The use of direct and continuous measurement to evaluate reading materials and procedures. Focus on Exceptional Children, 1970, 2, 1-11.

Mirkin, P. K., & Deno, S. L. Formative evaluation in the classroom: An approach to improving instruction (Research Report No. 10). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1979.

Mirkin, P. K., Deno, S. L., Tindal, G., & Kuehnle, K. Formative evaluation: Continued development of data utilization systems (Research Report No. 23), Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1980.

NaLDAP. Catalogue of child service demonstration centers: 1975-1976. Merrimac, Mass.: The Network, 1976.

NaLDAP. Title VI-G catalogue of child service demonstration centers: 1977-78. Merrimac, Mass.: The Network, 1978.

Nunnally, J. C. Tests and measurement: Assessment and prediction. New York: McGraw-Hill, 1959.

O'Connor, J. J., & Weiss, F. L. A brief discussion of the efficacy of raising standardized test scores by contingency reinforcement. Journal of Applied Behavior Analysis, 1974, 7, 351-352.

PDAS. Overview director and product guide: 1979-80. Seattle: University of Washington, 1980.

Pennypacker, H. S., Koenig, C. H., & Lindsley, O. R. Handbook of the standard behavior chart (preliminary ed.). Kansas City, Kan.: Precision Media, 1972.

Popham, W. J. Educational measurement for the improvement of instruction. Kappan, 1980, 61(8), 531-534.

Thorndike, E. L., & Lorge, I. The teacher's word book of 30,000 words. New York: Teachers College Press, 1944.

White, O. R. A pragmatic approach to the description of progress in the single case. Unpublished doctoral dissertation, University of Oregon, 1971.

White, O. R., & Haring, N. Exceptional teaching (2nd ed.). Columbus Ohio: Charles Merrill, 1980.

Table 1

Raw Score Means and Standard Deviations on the Fourteen Formative Evaluation Measures

	Resource[a]		Regular[b]		Combined[c]	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
PP-3 Isolated Words						
30-second	14.17	9.48	19.33	16.01	23.33	15.60
60-second	24.06	19.54	50.76	23.46	40.08	25.45
PP-6 Isolated Words						
30-second	8.97	7.19	19.41	10.02	15.23	10.30
60-second	19.14	14.39	39.46	20.67	31.33	20.83
3rd Grade Words in Context						
30-second	17.25	8.19	23.61	8.33	21.07	8.77
60-second	33.03	14.84	47.44	16.31	41.68	17.13
6th Grade Words in Context						
30-second	14.25	7.56	21.69	8.91	18.71	9.09
60-second	30.36	13.39	42.07	15.64	37.39	15.73
3rd Grade Oral Reading						
30-second	27.50	16.02	44.70	21.14	37.82	20.86
60-second	60.22	35.26	98.28	47.43	83.06	46.53
6th Grade Oral Reading						
30-second	23.53	13.95	40.93	20.74	33.97	20.09
60-second	52.28	28.26	84.69	41.75	71.72	39.95
Cloze	1.43	1.49	3.85	2.58	2.88	2.50
Word Meaning	5.33	2.10	6.77	3.43	6.19	3.03

[a] N=18.

[b] N=27.

Table 2

Correlation Matrix for Mean Correct Rate of Raw Scores on Fourteen Formative Evaluation Measures for the Resource Group. (N=18)[a]

	PP-3 Isolated Words		PP-6 Isolated Words		3rd Grade Words in Context		6th Grade Words in Context		3rd Grade Oral Reading		6th Grade Oral Reading			
	30 sec	60 sec	30 sec	60 sec	30 sec	60 sec	30 sec	60 sec	30 sec	60 sec	30 sec	60 sec	Cloze	Word Meaning
PP-3 Isolated Words														
30 sec		.97	.94	.97	.83	.83	.88	.89	.93	.91	.90	.93	.78	.66
60 sec			.95	.96	.77	.77	.87	.82	.88	.89	.88	.92	.75	.59
PP-6 Isolated Words														
30 sec				.96	.80	.82	.90	.87	.85	.89	.88	.90	.80	.67
60 sec					.83	.82	.92	.90	.90	.88	.93	.92	.84	.59
3rd Grade Words in Context														
30 sec						.92	.89	.93	.86	.86	.86	.89	.72	.80
60 sec							.91	.95	.87	.89	.87	.90	.82	.82
6th Grade Words in Context														
30 sec								.94	.88	.87	.92	.93	.84	.73
60 sec									.89	.89	.91	.92	.84	.75
3rd Grade Oral Reading														
30 sec										.96	.90	.95	.80	.66
60 sec											.88	.97	.78	.71
6th Grade Oral Reading														
30 sec												.92	.92	.60
60 sec													.80	.71
Cloze														.50
Word Meaning														

[a]All correlations are statistically significant ($p < .001$).

Table 3

Correlation Matrix for Mean Correct Rate of Raw Scores on Fourteen Formative Evaluation Measures for the Regular Group (N=27)[a]

	PP-3 Isolated Words		PP-6 Isolated Words		3rd Grade Words in Context		6th Grade Words in Context		3rd Grade Oral Reading		6th Grade Oral Reading			
	30 sec	60 sec	30 sec	60 sec	30 sec	60 sec	30 sec	60 sec	30 sec	60 sec	30 sec	60 sec	Cloze	Word Meaning
PP-3 Isolated Words														
30 sec		.95	.95	.94	.80	.83	.87	.84	.89	.91	.89	.92	.86	.59
60 sec			.97	.96	.86	.88	.92	.88	.91	.92	.89	.92	.84	.60
PP-6 Isolated Words														
30 sec				.97	.85	.87	.91	.88	.89	.91	.90	.92	.86	.62
60 sec					.86	.87	.92	.90	.89	.94	.90	.92	.86	.63
3rd Grade Words in Context														
30 sec						.96	.94	.95	.86	.87	.85	.87	.76	.71
60 sec							.95	.96	.87	.88	.86	.88	.81	.72
6th Grade Words in Context														
30 sec								.94	.90	.91	.90	.92	.85	.68
60 sec									.84	.87	.87	.86	.81	.75
3rd Grade Oral Reading														
30 sec										.97	.93	.96	.86	.49
60 sec											.95	.98	.86	.57
6th Grade Oral Reading														
30 sec												.97	.86	.55
60 sec													.87	.56
Cloze														.50
Word Meaning														

[a]All correlations are statistically significant ($p < .001$).

Table 4

Correlation Matrix for Mean Correct Rate of Raw Scores on Fourteen Formative Evaluation Measures for the Combined Group (N=45)[a]

	PP-3 Isolated Words		PP-6 Isolated Words		3rd Grade Words in Context		6th Grade Words in Context		3rd Grade Oral Reading		6th Grade Oral Reading			
	30 sec	60 sec	30 sec	60 sec	30 sec	60 sec	30 sec	60 sec	30 sec	60 sec	30 sec	60 sec	Cloze	Word Meaning
PP-3 Isolated Words														
30 sec	.95		.94	.94	.80	.83	.87	.84	.89	.91	.89	.92	.86	.59
60 sec			.96	.96	.86	.88	.92	.88	.91	.92	.89	.92	.84	.60
PP-6 Isolated Words														
30 sec				.97	.85	.87	.91	.88	.89	.91	.90	.92	.86	.62
60 sec					.86	.87	.92	.90	.89	.94	.90	.92	.86	.63
3rd Grade Words in Context														
30 sec						.96	.94	.95	.86	.87	.85	.87	.76	.71
60 sec							.95	.96	.87	.88	.86	.88	.81	.72
6th Grade Words in Context														
30 sec								.94	.90	.91	.90	.92	.85	.68
60 sec									.84	.87	.87	.86	.81	.75
3rd Grade Oral Reading														
30 sec										.97	.93	.96	.86	.49
60 sec											.95	.98	.86	.57
6th Grade Oral Reading														
30 sec												.97	.86	.55
60 sec													.87	.56
Cloze														.50
Word Meaning														

[a] All correlations are statistically significant ($p < .001$).

Table 5

Variability in Experimental Phases Expressed as Total Bounce and as Standard Error of the Estimate

	Total Bounce	Standard Error of Estimate
Student 1		
Phase A	8	2.71
Phase C	4	1.15
Average	6	1.93
Student 2		
Phase A	11	3.59
Phase C	12	4.24
Average	11.5	3.92
Grand Average		
Phases A & C, Students 1 & 2	8.75	2.92
Student 1		
Phase B	4	1.4
Phase D	0	.4
Average	2	.9
Student 2		
Phase B	4	1.75
Phase D	0	.44
Average	2	1.1
Grand Average		
Phases B & D, Students 1 & 2	2	.99

Table 6

Average Slope and Standard Error of Estimate (SEE)

for Three Sampling Domains

Domain Sampled	Slope	SEE
Grade-specific (GS)	.49	.29
Grade-entire (GE)	.20	.25
Across-grade (AG)	-.07	.29

Table 7

Comparisons of Sampling Domains on Slope and Standard Error of Estimate (SEE)

Paired Comparisons	Slope		SEE	
	t	p	t	p
GS with GE	4.05	.001	2.15	.05
GE with AG	2.68	.02	-.61	.55

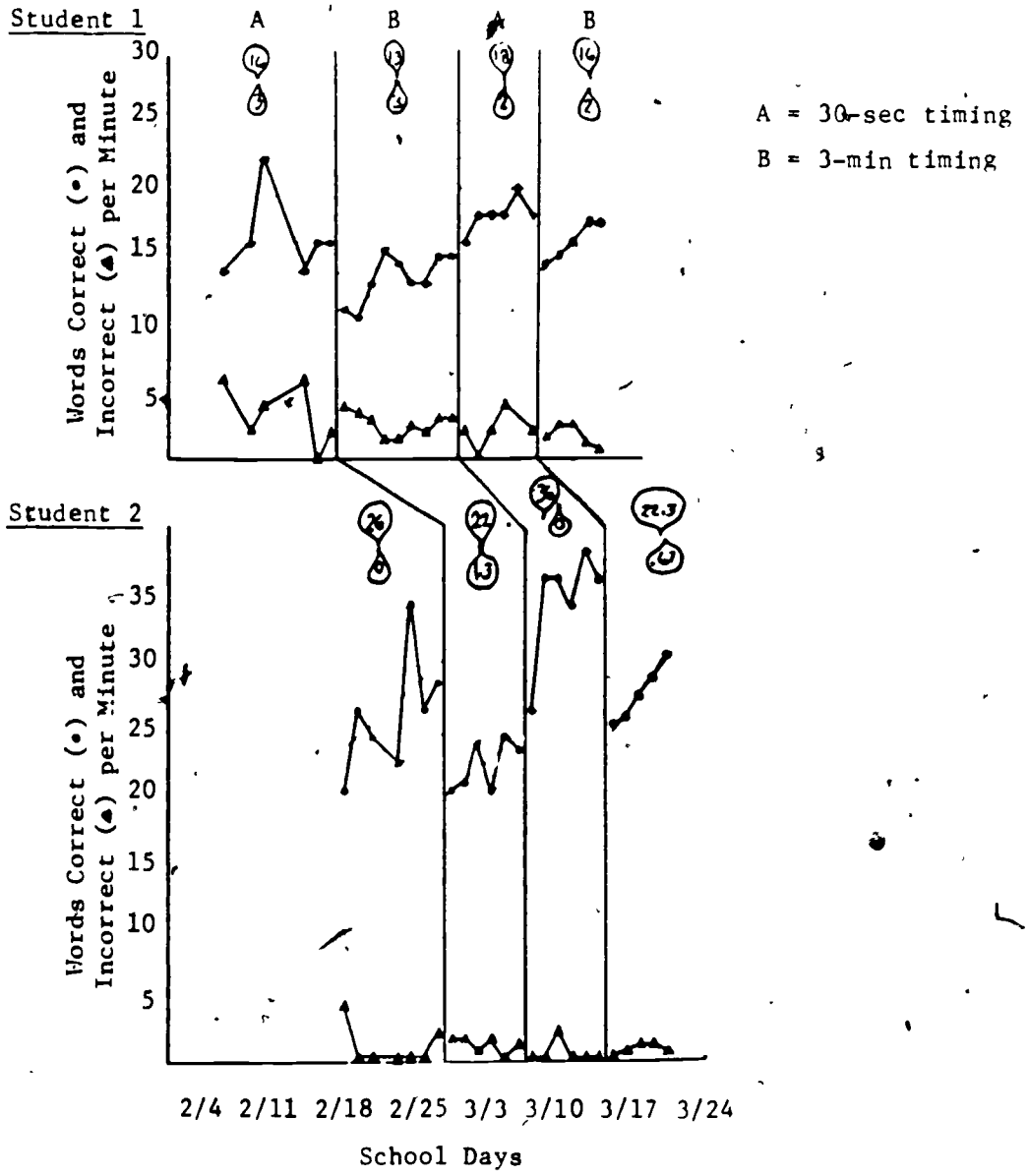

Figure 1. Words Correct and Words Incorrect per Minute for Students 1 and 2 during 30-sec and 3-min timings.

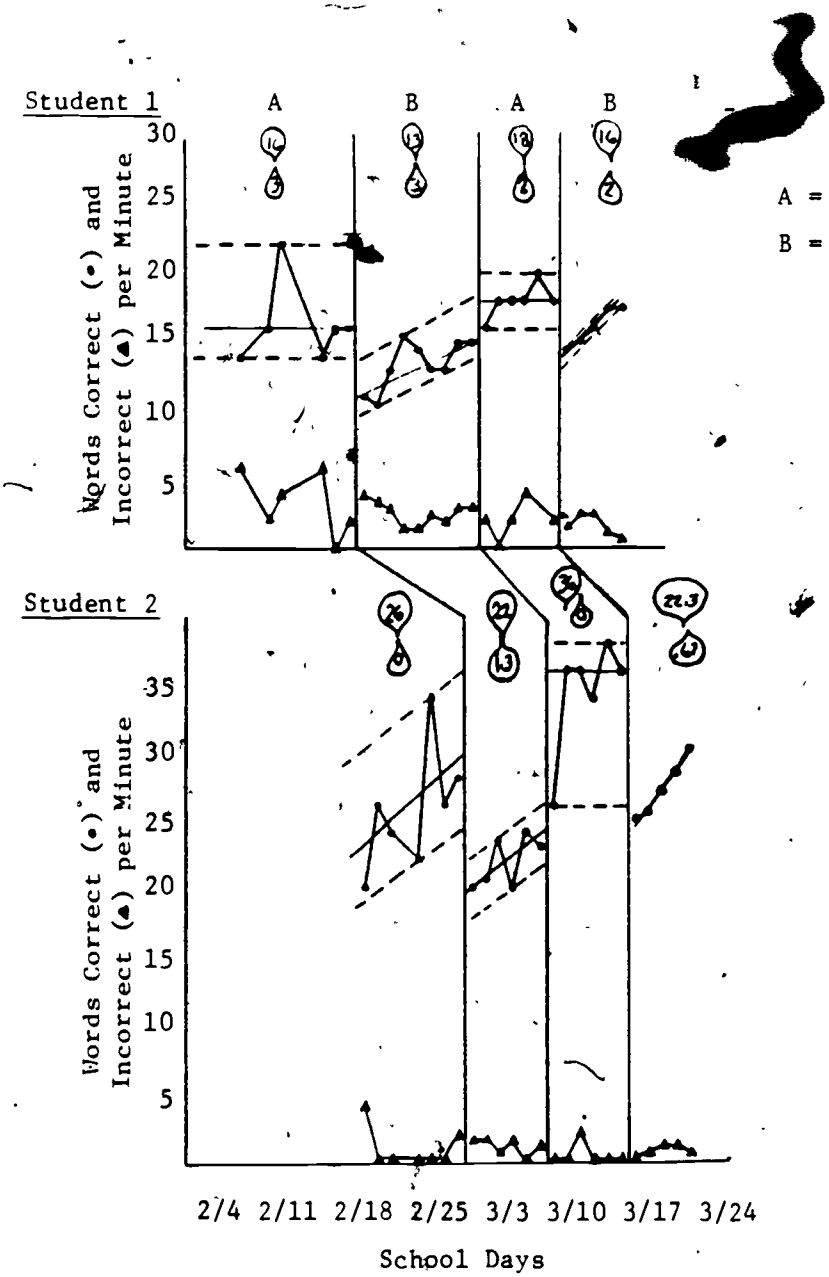

Figure 2. Total Bounce (TB) for Words Correct During Each Phase of the Experiment.

PUBLICATIONS

Institute for Research on Learning Disabilities
University of Minnesota

The Institute is not funded for the distribution of its publications. Publications may be obtained for $3.00 per document, a fee designed to cover printing and postage costs. Only checks and money orders payable to the University of Minnesota can be accepted. All orders must be prepaid.

Requests should be directed to: Editor, IRLD, 350 Elliott Hall; 75 East River Road, University of Minnesota, Minneapolis, MN 55455.

Ysseldyke, J. E. Assessing the learning disabled youngster: The state of the art (Research Report No. 1). November, 1977.

Ysseldyke, J. E., & Regan, R. R. Nondiscriminatory assessment and decision making (Monograph No. 7). February, 1979.

Foster, G., Algozzine, B., & Ysseldyke, J. Susceptibility to stereotypic bias (Research Report No. 3). March, 1979.

Algozzine, B. An analysis of the disturbingness and acceptability of behaviors as a function of diagnostic label (Research Report No. 4). March, 1979.

Algozzine, B., & McGraw, K. Diagnostic testing in mathematics: An extension of the PIAT? (Research Report No. 5). March, 1979.

Deno, S. L. A direct observation approach to measuring classroom behavior: Procedures and application (Research Report No. 6). April, 1979.

Ysseldyke, J. E., & Mirkin, P. K. Proceedings of the Minnesota roundtable conference on assessment of learning disabled children (Monograph No. 8). April, 1979.

Somwaru, J. P. A new approach to the assessment of learning disabilities (Monograph No. 9). April, 1979.

Algozzine, B., Forgnone, C., Mercer, C. D., & Trifiletti, J. J. Toward defining discrepancies for specific learning disabilities: An analysis and alternatives (Research Report No. 7). June, 1979.

Algozzine, B. The disturbing child: A validation report (Research Report No. 8). June, 1979.

Note: Monographs No. 1 - 6 and Research Report No. 2 are not available for distribution. These documents were part of the Institute's 1979-1980 continuation proposal, and/or are out of print.

Ysseldyke, J. E., Algozzine, B., Regan, R., & Potter, M. Technical adequacy of tests used by professionals in simulated decision making (Research Report No. 9). July, 1979.

Jenkins, J. R., Deno, S. L., & Mirkin, P. K. Measuring pupil progress toward the least restrictive environment (Monograph No. 10). August, 1979.

Mirkin, P. K., & Deno, S. L. Formative evaluation in the classroom: An approach to improving instruction (Research Report No. 10). August, 1979.

Thurlow, M. L., & Ysseldyke, J. E. Current assessment and decision-making practices in model programs for the learning disabled (Research Report No. 11). August, 1979.

Deno, S. L., Chiang, B., Tindal, G., & Blackburn, M. Experimental analysis of program components: An approach to research in CSDC's (Research Report No. 12). August, 1979.

Ysseldyke, J. E., Algozzine, B., Shinn, M., & McGue, M. Similarities and differences between underachieving and students labeled learning disabled: Identical twins with different mothers (Research Report No. 13). September, 1979.

Ysseldyke, J., & Algozzine, R. Perspectives on assessment of learning disabled students (Monograph No. 11). October, 1979.

Poland, S. F., Ysseldyke, J. E., Thurlow, M. L., & Mirkin, P. K. Current assessment and decision-making practices in school settings as reported by directors of special education (Research Report No. 14). November, 1979.

McGue, M., Shinn, M., & Ysseldyke, J. Validity of the Woodcock-Johnson psycho-educational battery with learning disabled students (Research Report No. 15). November, 1979.

Deno, S., Mirkin, P., & Shinn, M. Behavioral perspectives on the assessment of learning disabled children (Monograph No. 12). November, 1979.

Sutherland, J. H., Algozzine, B., Ysseldyke, J. E., & Young, S. What can I say after I say LD? (Research Report No. 16). December, 1979.

Deno, S. L., & Mirkin, P. K. Data-based IEP development: An approach to substantive compliance (Monograph No. 13). December, 1979.

Ysseldyke, J., Algozzine, B., Regan, R., & McGue, M. The influence of test scores and naturally-occurring pupil characteristics on psycho-educational decision making with children (Research Report No. 17). December, 1979.

Algozzine, B., & Ysseldyke, J. E. Decision makers' prediction of students' academic difficulties as a function of referral information (Research Report No. 18). December, 1979.

Ysseldyke, J. E., & Algozzine, B. Diagnostic classification decisions as a function of referral information (Research Report No. 19). January, 1980.

Deno, S. L., Mirkin, P. K., Chiang, B., & Lowry, L. Relationships among simple measures of reading and performance on standardized achievement tests (Research Report No. 20). January, 1980.

Deno, S. L., Mirkin, P. K., Lowry, L., & Kuehnle, K. Relationships among simple measures of spelling and performance on standardized achievement tests (Research Report No. 21). January, 1980.

Deno, S. L., Mirkin, P. K., & Marston, D. Relationships among simple measures of written expression and performance on standardized achievement tests (Research Report No. 22). January, 1980.

Mirkin, P. K., Deno, S. L., Tindal, G., & Kuehnle, K. Formative evaluation: Continued development of data utilization systems (Research Report No. 23). January, 1980.

Deno, S. L., Mirkin, P. K., Robinson, S., & Evans, P. Relationships among classroom observations of social adjustment and sociometric rating scales (Research Report No. 24). January, 1980.

Thurlow, M. L., & Ysseldyke, J. E. Factors influential on the psychoeducational decisions reached by teams of educators (Research Report No. 25). February, 1980.

Ysseldyke, J. E., & Algozzine, B. Diagnostic decision making in individuals susceptible to biasing information presented in the referral case folder (Research Report No. 26). March, 1980.

Thurlow, M. L., & Greener, J. W. Preliminary evidence on information considered useful in instructional planning (Research Report No. 27). March, 1980.

Ysseldyke, J. E., Regan, R. R., & Schwartz, S. Z. The use of technically adequate tests in psychoeducational decision making (Research Report No. 28). April, 1980.

Richey, L., Potter, M., & Ysseldyke, J. Teachers' expectations for the siblings of learning disabled and non-learning disabled students: A pilot study (Research Report No. 29). May, 1980.

Thurlow, M. L., & Ysseldyke, J. E. Instructional planning: Information collected by school psychologists vs. information considered useful by teachers (Research Report No. 30). June, 1980.

Algozzine, B., Webber, J., Campbell, M., Moore, S., & Gilliam, J. Classroom decision making as a function of diagnostic labels and perceived competence (Research Report No. 31). June, 1980.

Ysseldyke, J. E., Algozzine, B., Regan, R. R., Potter, M., Richey, L., & Thurlow, M. L. Psychoeducational assessment and decision making: A computer-simulated investigation (Research Report No. 32). July, 1980.

Ysseldyke, J. E., Algozzine, B., Regan, R. R., Potter, M., & Richey, L. Psychoeducational assessment and decision making: Individual case studies (Research Report No. 33). July, 1980.

Ysseldyke, J. E., Algozzine, B., Regan, R., Potter, M., & Richey, L. Technical supplement for computer-simulated investigations of the psychoeducational assessment and decision-making process (Research Report No. 34). July, 1980.

Algozzine, B., Stevens, L., Costello, C., Beattie, J., & Schmid, R. Classroom perspectives of LD and other special education teachers (Research Report No. 35). July, 1980.

Algozzine, B., Siders, J., Siders, J., & Beattie, J. Using assessment information to plan reading instructional programs: Error analysis and word attack skills (Monograph No. 14). July, 1980.

Ysseldyke, J., Shinn, M., & Epps, S. A comparison of the WISC-R and the Woodcock-Johnson Tests of Cognitive Ability (Research Report No. 36). July, 1980.

Algozzine, B., & Ysseldyke, J. E. An analysis of difference score reliabilities on three measures with a sample of low achieving youngsters (Research Report No. 37). August, 1980.

Shinn, M., Algozzine, B., Marston, D., & Ysseldyke, J. A theoretical analysis of the performance of learning disabled students on the Woodcock-Johnson Psycho-Educational Battery (Research Report No. 38). August, 1980.

Richey, L. S., Ysseldyke, J., Potter, M., Regan, R. R., & Greener, J. Teachers' attitudes and expectations for siblings of learning disabled children (Research Report No. 39). August, 1980.

Ysseldyke, J. E., Algozzine, B., & Thurlow, M. L. (Eds.). A naturalistic investigation of special education team meetings (Research Report No. 40). August, 1980.

Meyers, B., Meyers, J., & Deno, S. Formative evaluation and teacher decision making: A follow-up investigation (Research Report No. 41). September, 1980.

Fuchs, D., Garwick, D. R., Featherstone, N., & Fuchs, L. S. On the determinants and prediction of handicapped children's differential test performance with familiar and unfamiliar examiners (Research Report No. 42). September, 1980.

Algozzine, B., & Stoller, L. Effects of labels and competence on teachers' attributions for a student (Research Report No. 43). September, 1980.

Ysseldyke, J. E., & Thurlow, M. L. (Eds.). The special education assessment and decision-making process: Seven case studies (Research Report No. 44). September, 1980.

Ysseldyke, J. E., Algozzine, B., Potter, M., & Regan, A. A descriptive study of students enrolled in a program for the severely learning disabled (Research Report No. 45). September, 1980.

Marston, D. Analysis of subtest scatter on the tests of cognitive ability from the Woodcock-Johnson Psycho-Educational Battery (Research Report No. 46). October, 1980.

Algozzine, B., Ysseldyke, J. E., & Shinn, M. Identifying children with learning disabilities: When is a discrepancy severe? (Research Report No. 47). November, 1980.

Fuchs, L., Tindal, J., & Deno, S. Effects of varying item domain and sample duration on technical characteristics of daily measures in reading (Research Report No. 48). January, 1981